GenePattern

GeneNeighbors Documentation

Module name:	GeneNeighbors
Description:	Select genes that most closely resemble a continuous profile (e.g., another gene)
Author:	Ken Ross (Broad Institute), gp-help@broad.mit.edu
Date:	5/2/05
Release:	2

Summary: The GeneNeighbor marker analysis algorithm calculates the nearest neighbors for a particular gene (or other continuous vector pseudo-gene) by trying to find other genes whose expression values follow similar trends for the samples. The user specifies the number of nearest neighbors to find for a particular gene by entering a value for the Num.Neighbors parameter (defaults to 50). There are four choices for the distance metric: Cosine (default), Euclidean, Manhattan, and Pearson. The cosine distance is given by $d_C = \sum_i x_i * y_i \Big/ \left(\sum_i x_i^2 * \sum_i y_i^2 \right)^{1/2}$ where i is the sample number, x_i is the named reference gene's expression value for sample i, and y_i is the expression value of the gene we are testing. The Euclidean distance is given by $d_E = \left(\sum_i (x_i - y_i)^2 \right)^{1/2}$ where i is the sample number, x_i is the named reference gene's expression value, and y_i is the expression value of the gene we are testing. The Manhattan distance is given by $d_M = \sum_i |x_i - y_i|$ where i is the sample number, x_i is the named reference gene's expression value, and y_i is the expression value of the gene we are testing. The Pearson distance is calculated by $d_P = n\left(\sum_i x_i * y_i \right) - \left(\sum_i x_i \right) * \left(\sum_i y_i \right) \Big/ \left(\left[n \sum_i x_i^2 - \left(\sum_i x_i \right)^2 \right] * \left[n \sum_i y_i^2 - \left(\sum_i y_i \right)^2 \right] \right)^{1/2}$ where i is the sample number, n is the number of samples, x_i is the named reference gene's expression value, and y_i is the expression value of the gene we are testing. Running this algorithm produces a table with three columns: 1) Feature - contains the gene's identifier from the input file, 2) Desc - contains the gene's description from the input file, and 3) Score - contains the calculated distance for the gene relative to the reference gene. The genes in the output table are ordered based upon the score.

Our implementation of the GeneNeighbors algorithm also includes several basic data pre-processing options. The thresholding option allows the user to set minimum and maximum thresholds for the data. Any value in the data set that is less then the value for the minimum threshold is set to the minimum threshold value. Similarly, any value that is greater than the maximum threshold is set to the maximum threshold value. There is also a variation filter that will remove rows from the data set whose values do not vary greatly. For a given row (gene), minVal is the minimum value in that row and maxVal is the maximum value in that row. If maxVal/minVal is greater than the specified minimum fold difference variation ratio (defaults to 5) and maxVal - minVal is greater than the specified minimum absolute difference (defaults to 50), then the row passes the filter. Any rows that do not pass the filter are excluded from the list of features that can be used for the GeneNeighbors gene ranking algorithm. If the

reference gene is filtered out or more neighbors are requested than the number of genes remaining after filtering, an error will be produced.

The results table from the GeneNeighbors algorithm can be viewed with the GeneListSignificanceViewer and the data results file can be viewed with the HeatMapViewer.

References:

- Golub T.R., Slonim D.K., et al. “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” Science, 531-537 (1999). and the supplemental information on the website http://www-genome.wi.mit.edu/cgi-bin/cancer/publications/pub_menu.cgi for a more complete description of marker permutation testing.

- Slonim, D.K., Tamayo, P., Mesirov, J.P., Golub, T.R., Lander, E.S. (2000) Class prediction and discovery using gene expression data. In Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB) 2000. ACM Press, New York, pp. 263–272.

Parameters:

Name	Description	Choices
data.filename:	data file (gct or res)	
gene.accession:	reference gene accession from data file to find neighbors for	
num.neighbors:	number of neighbors to find	
marker.list.file:	output filename for analysis results (.odf format)	
marker.dataset.file:	output filename raw data for selected markers (.gct format)	
distance.metric:	continuous metric for finding neighbors	0=Cosine distance (default), 1=Euclidean distance, 2=Manhattan distance, 3=Pearson
filter.data	if no, values below will be ignored	yes/no
min.threshold:	minimum threshold for data	
max.threshold:	maximum threshold for data	
min.fold.diff:	minimum fold difference for filtering genes	default value: 5
min.abs.diff:	minimum absolute difference for filtering genes	

Return Value:

1. marker.list.file: output file (.odf format) with table of analysis results.
2. marker.dataset.file: output file (gct format) with raw data for selected markers.

Platform dependencies:

Task type:	GeneListSelection
CPU type:	any
OS:	any
Java JVM level:	1.4
Language:	Java